MICROCOPY RESOLUTION TEST CHART

BR98722 ②

AD-A165 661

RSRE
MEMORANDUM No. 3897

ROYAL SIGNALS & RADAR ESTABLISHMENT

ON EXTENSIONS TO FISHERS LINEAR DISCRIMINANT
FUNCTION

Author: I D Longstaff

RSRE MEMORANDUM No. 3897

DTIC FILE COPY

PROCUREMENT EXECUTIVE,
MINISTRY OF DEFENCE,
R S R E MALVERN,
WORCS.

DTIC
SELECTED
MAR 2 4 1986
S
E

86 3 21 056

ROYAL SIGNALS AND RADAR ESTABLISHMENT

Memorandum 3897

TITLE: ON EXTENSIONS TO FISHER'S LINEAR DISCRIMINANT FUNCTION

AUTHOR: I D Longstaff

DATE: November 1985

SUMMARY

     This Memorandum describes how Fisher's Linear Discriminant can
be combined with the Fukunaga-Koontz transform to give a useful
technique for reduction of feature space from many to two or three
dimensions. Performance is seen to be superior in general to the
Foley-Sammon extension to Fisher's method. The technique is then
extended to show how a new radius vector (or pair of radius vectors)
can be combined with Fisher's vector to produce a classifier with
even more power of discrimination. Illustrations of the technique
show that good discrimination can be obtained even if there is
considerable overlap of classes in any single projection.

     Index Terms ← Dimensionality reduction, discriminant vectors,
feature selection, Fisher criterion, linear transformations,
separability.

Accession For

NTIS GRA&I	X
DTIC TAB	☐
Unannounced	☐
Justification	

By _____
Distribution/

Availability Codes

Dist | Avail and/or Special

A-1

RSRE MEMORANDUM NO 3897

ON EXTENSIONS TO FISHER'S LINEAR DISCRIMINANT FUNCTION

I D Longstaff

LIST OF CONTENTS

LIST OF FIGURES

LIST OF APPENDICES

1   INTRODUCTION

Fisher's linear discriminant function[1,2] makes a useful classifier where the two classes have features with well separated means compared with their scatter. The method finds that vector which, when training data is projected on to it, maximises the class separation. It is a many to one linear transformation.

At the other extreme, for the case in which both classes have the same mean but different variances Fukanaga and Koontz have described[4] a transform which maximises class discrimination. This method transforms the data so the joint (sum) covariance matrix for the two classes is the identity matrix and then selects the eigenvector with the largest difference in eigenvalues for the two classes. The data is then projected on to this eigenvector to produce the classifier.

1

An advantage of linear projections such as these is that they can give the system designer some appreciation of class separation if training data is presented as histograms of points projected on to the discriminant vector. An even better appreciation is gained if training data is presented as a two-dimensional scatter diagram where the dimensions are chosen to be those which show the best, in some sense, classification. Such projections on to two dimensions also allows the operator to select curved or piecewise linear decision boundaries as a pattern classifier. This interactive approach to classifier design has been found to yield more productive results than the use of non-interactive methods such as the quadratic discriminant function especially where data is non-Gaussian.

The quadratic discriminant method operates thus: first the training data is used to estimate the class means and covariance matrices. These are then used to generate multivariate Gaussian probability distribution functions for the two classes. The decision as to which class a new data point belongs is taken by evaluating the probability functions at that point and the new point is assigned to that class with the greatest probability. If the data is truly Gaussian and if enough training points are available to estimate the means and covariances accurately this method produces the optimum Baysian classifier. The decision boundary consists of those points where the pfd's are equal and this will in general be a multidimensional quadratic surface in feature space.

In practice training data is often non-Gaussian and the interactive approach using projections of the data is preferred. The problem is that there is often a huge number of combinations of pairs of features which can be examined and a methodology is needed which standardises the data and points to possible two-dimensional projections where discrimination may be high. Ideally the method should also project the Baysian decision boundary into a unique line in the two-dimensional subspace. This will then maintain the performance of the classifier to the Baysian rate if the data should happen to be Gaussian.

Foley and Sammon[3] have suggested an extension to Fisher's method which gives a two-dimensional (or more) projection for displaying data. Their method is based on finding Fisher's vector first; the data is then projected on to the subspace normal to Fisher's vector and the process of finding Fisher's vector in that subspace is repeated. The data is then displayed projected on to the plane subtended by these two vectors and a decision boundary is constructed in that plane. It is shown later in this Memorandum that this method is of doubtful value for finding the best classification subspace.

The methods proposed in this Memorandum are based on applying a standardising transform to the training data. This then allows Fisher's method to be used in conjunction with Fukunaga's method to select the best two-dimensional linear projection. The method is then extended to show how a nonlinear combination of features can result in a two or three dimensional scatter diagram with a performance which is round to be better than the linear method in a number of cases. The method further allows the Baysian decision surface to be uniquely represented by a line in the subspace for multivariance Gaussian data with certain conditions.

2 OBSERVATIONS ON FISHER'S METHOD

Fisher's method finds the vector $\underline{F}$ which gives greatest (as defined by a criterion function) class separation to data points projected on to the vector. The criterion function is:

$$J(\underline{F}) = \frac{|\underline{F}^T (\underline{\mu}_a - \underline{\mu}_b)|^2}{\underline{F}^T (W_a + W_b)\underline{F}}$$

where $\underline{\mu}_i$ = class mean for class i, i = a, b

W_i = covariance matrix for class i.

The vector solution to this maximising problem can be shown to be:

$$\underline{F} = [W_a + W_b]^{-1} [\underline{\mu}_a - \underline{\mu}_b]$$

It should be mentioned that maximising this criterion function does not neces-sarily produce the best projection for classification as shown by Malina[5]. However the differences are usually very small and Fisher's method is used in this Memorandum because it leads to the interesting generalisations and extensions shown here.

It is clear that a data set can be transformed on to a new set of co-ordinates without loss or gain of discriminating performance provided the transform is unique (ie invertible). A decision boundary in one co-ordinate system maps on to the other with the same number of true of false classifica-tions on either side. Now the decision threshold on the Fisher axis corres-ponds to a hyperplane decision surface in feature space, where the hyperplane is normal to the Fisher axis and intersecting it at the decision threshold. It is shown in Appendix A that if a linear transformation is applied to the data the same decision threshold is generated if the Fisher vector is found either before or after the transformation.

An interesting transform which can be applied to a training data set is that which causes the joint scatter matrix for the training data to become an identity matrix. Such a transform can be visualised by first applying a rota-tion of axes (orthogonal transform) so that the eigenvectors of the scatter matrix are the orthogonal co-ordinate system (the Karhunen-Loeve transform). Each co-ordinate can then be scaled so that the variances are unity thus giving a unity scatter matrix.

In this new co-ordinate system the Fisher axis is

$$\underline{F}' = I (\underline{\mu}_a' - \underline{\mu}_b') \ .$$

That is, F' is parallel to the axis intercepting the means of the two distribu-tions, see Figure 1 and 2. This appears to be a useful way of standardising the use of Fisher's method and there is no loss or gain in performance of the transformed training data compared with the method applied to the untransformed data.

It is also evident that if we apply this standardising transform and then project the data on to the hyperplane normal to the new Fisher vector then the two distributions obtained will have coincident means (see Figure 3).

3

3   EXTENSIONS TO FISHER'S METHOD

As mentioned earlier it would be convenient if we could combine Fisher's vector with some other discriminant function to give a two-dimensional vector representation - simply because it is easy to plot two-dimensional scatter diagrams, and also two dimensions should, in general, give better discrimination than one dimension.

So on what basis do we select another dimension? If the standardising transform is applied first then, as we have seen, the clusters in the subspace normal to Fisher's axis will have coincident means. This makes the task of finding a second Fisher axis impossible. It is this fact that makes this method rather suspect and it was recognition of this which lead to the identification of the generalisations described in this Memorandum. It is believed that these new methods do improve discrimination and indeed illustrative examples are given to show the improvements which can be obtained when the methods are used instead of Foley-Sammon.

4   FISHER WITH FUKUNAGA-KOONTZ

In all the methods described from here onwards the first step is to transform the data to give an identity joint covariance matrix (the standardising transform). The Fisher axis is then the axis through the means. If the data is projected on to the hyperplane perpendicular to the Fisher axis then the means will be coincident. To obtain maximum difference between the two classes we can look for the projection which maximises the differences in variances (normalised by the sum of the variances). It is then evident from Figure 4 that the bigger the difference the better the classifier.

If the two classes have covariances W_a and W_b. Let T be the standardising transform such that:

$$T (W_a + W_b) T^{-1}$$

Fukunaga showed that the eigenvectors of

$$T W_a T^{-1} \quad \text{and} \quad T W_b T^{-1}$$

are the same and that all eigenvalues are bounded by 0 and 1 and that the sum of any pair equals 1, ie $\lambda_{ia} + \lambda_{ib} = 1$.

It is clear from this that the axis which gives the biggest difference in variances for the two classes is the eigenvector with the biggest difference in eigenvalue for the two classes.

Thus the Fisher projection with the Fukunaga-Koontz (F-K) projection gives a many-to-one transform with a performance usually better and never worse than the Fisher with the Foley-Sammon (F-S) projection (for multivariate Gaussian data). Figures 6 and 7 illustrate an example where the two classes have different means and where the F-S and F-K vectors are different. Figure 5 shows the parameters used to generate the test data.

Figure 6 shows the scatter diagram for F-S with 100 points for each class to train and test. Figure 7 shows the same data with Fisher and F-K indicating a clear improvement.

4

5 FISHER WITH A RADIUS VECTOR

This use of the F-K transform suggests an even more powerful (nonlinear) many-to-one transform. Consider a three feature two class problem. After the standardising transform and projection on to the plane normal to the Fisher axis we might obtain distributions as shown in Figure 3. If all the eigenvalues for Class A are less than those for Class B then the Baysian decision surface can be shown to be an ellipse with the eigenvectors as axes (see Appendix B).

By rotating the data and rescaling it, it is clear that the Baysian surface can be made into a circle with Class A inside and Class B outside. The only information needed to test if a new data point lies inside or outside is to compute its radius from the common mean and test against the radius to the Baysian threshold. In the more general multidimensional case the Baysian surface can be made into one hypersphere if the eigenvalues of one class are all less than those of the other: classification in this case involves assigning a new point to Class A or B according to whether it lies inside (class with smaller eigenvalues) or outside the hypersphere (class with larger eigenvalues).

For either of these cases the data set can be mapped from the original multidimensional feature space down to two where distance along the Fisher axis is one feature and radius (Euclidean distance) from the Fisher axis (in the transformed space) is the other feature.

This method will be better than the use of Fisher with one F-K axis alone. In circumstances where the class distribution functions have circular symmetry along the Fisher axis the Baysian surface will also have circular symmetry and map on to a unique line in the Radius-Fisher plane. Hence in this case performance of the F-R space is optimal and equal to performance of a Baysian classifier in full feature space (for Gaussian data).

To use this method in the more general case where not all eigenvalues of one class are less than those of the other we divide the training data into two subsets of reduced dimensionality where the first subspace only contains features where $\lambda_{ia} < \lambda_{ib}$ and the second only those features where $\lambda_{ib} < \lambda_{ia}$. Any features for which should not be included in the radius calculation. Figure 11 shows a schematic diagram of the classifier using this method. If a two-dimensional classifier is required the subset with the best performance can be selected, otherwise a three-dimensional classifier can be constructed.

Figure 8 shows the scatter in the F-R axis using the same data as used for the Fisher F-S method (Figure 5). It is seen that the error rate reduces from 6% with F-K to 2% with the radius for this example. The advantage of using the radius vector can be seen even more clearly when more features are available as in the ten-dimensional example of Figure 9. Notice that there is no increase in mean differences in variances. The scatter diagram obtained using the Fisher-Radius method indicates an average error rate approaching zero. If the linear methods are applied they would show no improvement over the five-dimensional case because no use is made of the additional features. All the data shown here was generated to give multivariate Gaussian statistics.

The quadratic discriminant function gives optimum performance if data is multivariate Gaussian but the additional scope given by the procedures described hence allow a better performance to be obtained if data is highly non-Gaussian. For instance we found that the data generated with a negative exponential distribution was better classified using the Fisher radius vector than using the quadratic discriminant function.

6   CONCLUSIONS

If data is standardised with a linear transform to give a unit joint covariance matrix the Foley-Sammon axis becomes meaningless because the two classes have coincident means in the hyperplane normal to the Fisher axis. In this case the Fukunaga-Koontz transform allows the next best feature to be selected.

Another simple linear transform can give a spherical Baysian decision surface on features where all eigenvalues are smaller for one of the classes (see Appendix B). In this case distance along the Fisher axis and radius from the Fisher axis form a powerful discriminating function. If all eigenvalues are not smaller for one class then the features can be divided into two groups and two radii calculated with the best or both being used with Fisher distance to provide the discriminating function. Both linear transforms can be combined into a single operation. Figure 11 shows how simple the implementation of this classifier would be.

The arguments used in this paper apply to multivariate Gaussian distributions. In practice distributions are not so simple. However we believe that data can be standardised and inspected using the procedures described here as a first approach to the classifier design problem. A good classifier may result, perhaps with some exercising of pathological features or with the inclusion of special stages to include highly non-Gaussian but well discriminating features.

The methods described here are for two-class problems only. However they are particularly suited to the technique of reducing a many-class problem to that of many two-class problems.

In this form the problem is to identify one species against the world background of other species. This usually results in the world background class having a larger variance compared with the required species and our method takes advantage of this characteristic and can give good performance even when the two class means are similar.

REFERENCES

1   Fisher, R A; "The use of multiple measurements in taxonomic problems", Ann. Eugenica, 7, Part II, pp 179-188, 1936.

2   Duda, R and Hart, P; Pattern Classification and Scene Analysis, New York: Wiley, 1973, pp 114-129, 221-225.

3   Foley, D and Sammon, J; "An optimal set of discriminant vectors", IEEE Trans. Comput., Vol. C-26, pp 281-289, March 1975.

4   Fukunaga, K and Koontz, W L; "Application of the Karhunen-Loeve expansion to feature selection and ordering", IEEE Trans. Comput., Vol. C-19, pp 311-318, April 1980.

5   Malina, W; "On an extended Fisher criterion for feature selection",
IEEE Trans. on pattern An., Vol. PAMI-3, pp 611-614, September 1981.

APPENDIX A   INVARIANCE OF FISHER'S METHOD

To show that the decision surface produced by the Fisher linear discriminant function is invariant under a transform T where $\underline{y} = T \underline{x}$ is a non-singular linear transform.

Suppose we have a data set $\underline{x}$ (where $\underline{x}$ is an L-dimensional vector). The mean and covariance of $\underline{x}$ are then defined as

$$\underline{\mu} = E(\underline{x})$$

$$W = E(\underline{x}\,\underline{x}^T)$$

The Fisher discriminant vector $\underline{F}$ for the two classes is then found from[1]

$$\underline{F} = (W_a + W_b)^{-1} (\underline{\mu}_a - \underline{\mu}_b) \qquad (A1)$$

Suppose data is now transformed by T such that $\underline{y} = T\underline{x}$ where T is a non-singular linear transform. Then the covariance matrix of the transformed data is:

$$W' = E(T\underline{x}\,\underline{x}^T\,T^T) = T W T^T$$

Let the vector $\underline{p}$ be normal to the Fisher vector in the untransformed space, ie

$$\underline{p}^T\underline{F} = 0 \qquad (A2)$$

The Fisher vector $\underline{F}'$ in the transformed space is obtained from:

$$
\begin{aligned}
\underline{F}' &= (W_a' + W_b')^{-1} T(\underline{\mu}_a - \underline{\mu}_b) \\
&= (TW_aT^T + TW_bT^T)^{-1} T(\underline{\mu}_a - \underline{\mu}_b) \\
&= (T(W_a + W_b)T^T)^{-1} T(\underline{\mu}_a - \underline{\mu}_b) \\
&= (T^T)^{-1} (W_a + W_b)^{-1} T^{-1} T(\underline{\mu}_a - \underline{\mu}_b) \\
&= (T^T)^{-1} (W_a + W_b)^{-1} (\underline{\mu}_a - \underline{\mu}_b) \qquad (A3)
\end{aligned}
$$

We can show that any plane normal to the Fisher discriminant vector in untransformed space will be normal to the new Fisher vector in transformed space if we can show

$$\underline{p}'^T \underline{F}' = 0 \quad .$$

But

$$
\begin{aligned}
\underline{p}'^T \underline{F}' &= (T\underline{p})^T (T^T)^{-1} (W_a + W_b)^{-1} (\underline{\mu}_a - \underline{\mu}_b) && \text{from (A3)} \\
&= \underline{p}^T T^T (T^T)^{-1} (W_a + W_b)^{-1} (\underline{\mu}_a - \underline{\mu}_b) \\
&= \underline{p}^T (W_a + W_b) (\underline{\mu}_a - \underline{\mu}_b) \\
&= \underline{p}^T\underline{F} && \text{from (A1)} \\
&= 0 && \text{from (A2)}
\end{aligned}
$$

APPENDIX B   HYPERSPHERICAL DECISION BOUNDARIES

This Appendix shows how and when the Baysian decision surface on multi-variate Gaussian distributions with common means can be transformed to a hypersphere.

The Fukunaga-Koontz transform shows how two distributions can be transformed to have common eigenvectors. The K-L transform extracted from the covariance matrix of one class can be used to align the eigenvectors with a new co-ordinate system. In this system the two data sets are decorrelated and as they have common means the pdf's can be written as:

Class A

$$P(a) = \prod_{i=1}^{L} \left(\frac{1}{(2\pi)^{L/2} \lambda_{ia}^{\frac{1}{2}}} \exp - x_i^2/2\lambda_{ia} \right) \tag{B1}$$

Class B

$$P(b) = \prod_{i=1}^{L} \left(\frac{1}{(2\pi)^{L/2} \lambda_{ib}^{\frac{1}{2}}} \exp - x_i^2/2\lambda_{ib} \right) \tag{B2}$$

At the Baysian decision boundary the pdf's are equal. Taking logs of equations (B1) and (B2) and equating gives:

$$\sum_{i=1}^{L} x_i^2/2\lambda_{ia} + \log \prod_{i=1}^{L} \lambda_{ia}^{\frac{1}{2}} = \sum_{i=1}^{L} x_i^2/2\lambda_{ib} + \log \prod_{i=1}^{L} \lambda_{ib}^{\frac{1}{2}}$$

or

$$\sum_{i=1}^{L} x_i^2 (1/2\lambda_{ia} - 1/2\lambda_{ib}) = \frac{1}{2} \log \prod_{i=1}^{L} \left(\frac{\lambda_{ib}}{\lambda_{ia}} \right) \tag{B3}$$

If all the coefficients of x have the same sign this is the equation of a hyperellipse. A simple rescaling transform can be used to reduce this to a hypersphere with the scaling factor in the ith co-ordinate being given by

$$k_i^2 = 1/\lambda_{ia} - 1/\lambda_{ib}$$

and the actual transform being

$$\underline{x}' = \begin{bmatrix} k_1 & & & & 0 \\ & k_2 & & & \\ & & \cdot & & \\ & & & \cdot & \\ & & & & \cdot \\ 0 & & & & k_L \end{bmatrix} \underline{x}$$

If the coefficients of x in equation (3) have different signs some sections of the surface will be a saddle. To overcome this the data can be transformed into two subsets of similarly signed coefficients. If the variances of two features are equal

$$\lambda_{ia} = \lambda_{ib}$$

then the ith feature will give no further discrimination and can be disregarded.
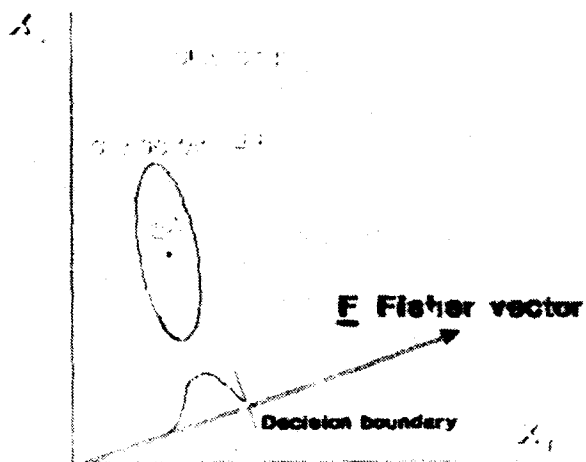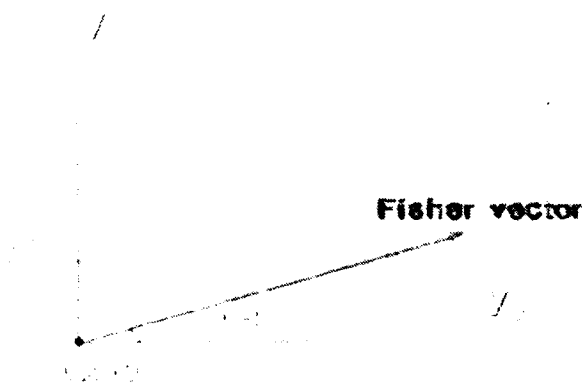
F Fisher vector

Decision boundary

Fig 1   Fishers Linear Discriminant Function

Fisher vector

Fig 2   Fishers Linear Discriminant Function
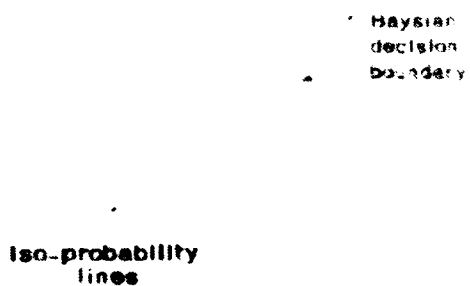after applying transform to make
$(W_a + W_b) = (I)$ & $\mu_a = 0$

Baysian
decision
boundary

Fig 3   Projection of Data onto subspace
normal to F

Iso-probability
lines
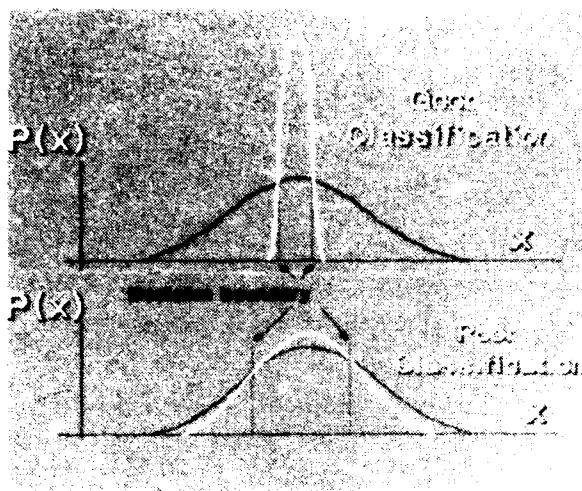
Fig 4   Classification of data with common mean and
       (a)   big difference in variances
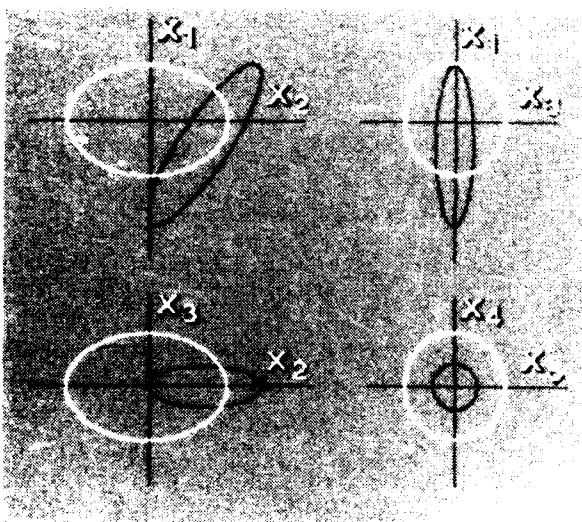   (b)   small difference in variances

Fig 5   Some projections of the data used in Fig 6-8 with:

$$\mu_a = 0 \quad \mu_b = (-\tfrac{1}{2}, \sqrt{3}, \ 0, \ 0, \ 0)^T$$

$$(W_a) = \begin{bmatrix} 1 & & & & \\ & 4 & & 0 & \\ & & 1 & & \\ & 0 & & 1 & \\ & & & & 1 \end{bmatrix}$$

$$(W_b) = \begin{bmatrix} 3.02 & 3.41 & 0 & 0 & 0 \\ 3.41 & 4.12 & 0 & 0 & 0 \\ 0 & 0 & .0625 & 0 & 0 \\ 0 & 0 & 0 & .25 & 0 \\ 0 & 0 & 0 & 0 & .25 \end{bmatrix}$$
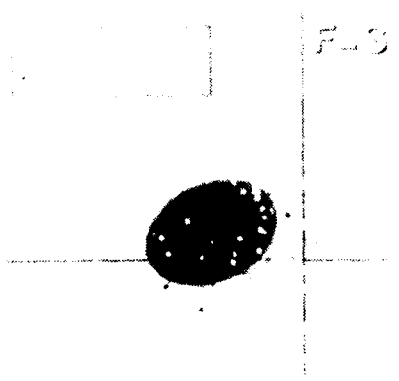

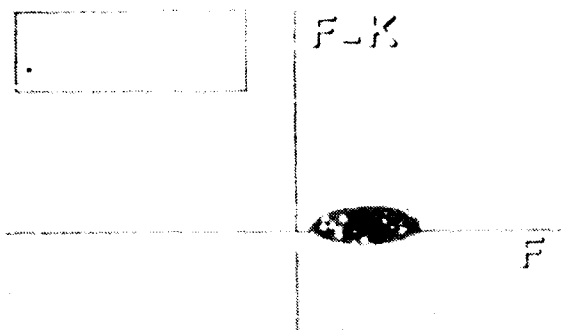Fig 6   Scatter in Fisher Foley-Sammon plane
Total errors $\sim$ 12%

F−K

F

Fig 7   Scatter in Fisher Fukanaga-Koontz plane
Total errors ∼ 6.5%

R

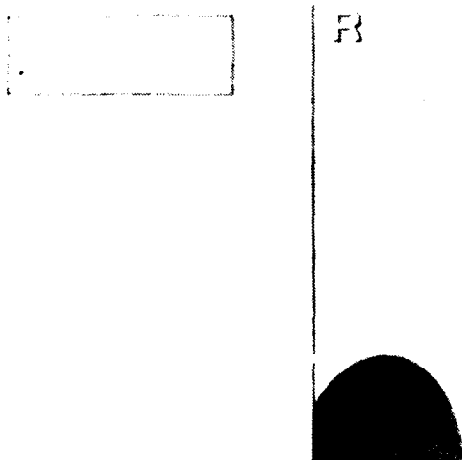Fig 8   Scatter in Fisher Radius plane
Total errors ∼ 2%

Fig 9   Scatter in Fisher Radius Plane,
10 dimension data generated with
$\mu_a = 0 \quad \mu_b = (-\tfrac{1}{2}\sqrt{3} \quad 0 \quad 0 \ldots \ldots 0)^T$

R

F

$$W_a \begin{bmatrix} 1 & & & & \\ & 4 & & 0 & \\ & & 1 & & \\ & 0 & 1 & & \\ & & & \ddots & \\ & & & & \ddots \\ & & & & & 1 \end{bmatrix}$$

$$W_b \begin{bmatrix} 3.02 & 3.41 & 0 & & \\ 3.41 & 4.12 & 0 & & \\ 0 & 0 & .0625 & & \\ & & & \ddots & \\ & & & & .0625 \end{bmatrix}$$

N features
in
↓

Fisher distance

NxN Trans-form

sq
sq
sq

$\oplus$ (Radius 1)^2

$(\lambda_a < \lambda_b)$

sq
sq

$\oplus$ (Radius 2)^2

$(\lambda_a > \lambda_b)$

not used
if
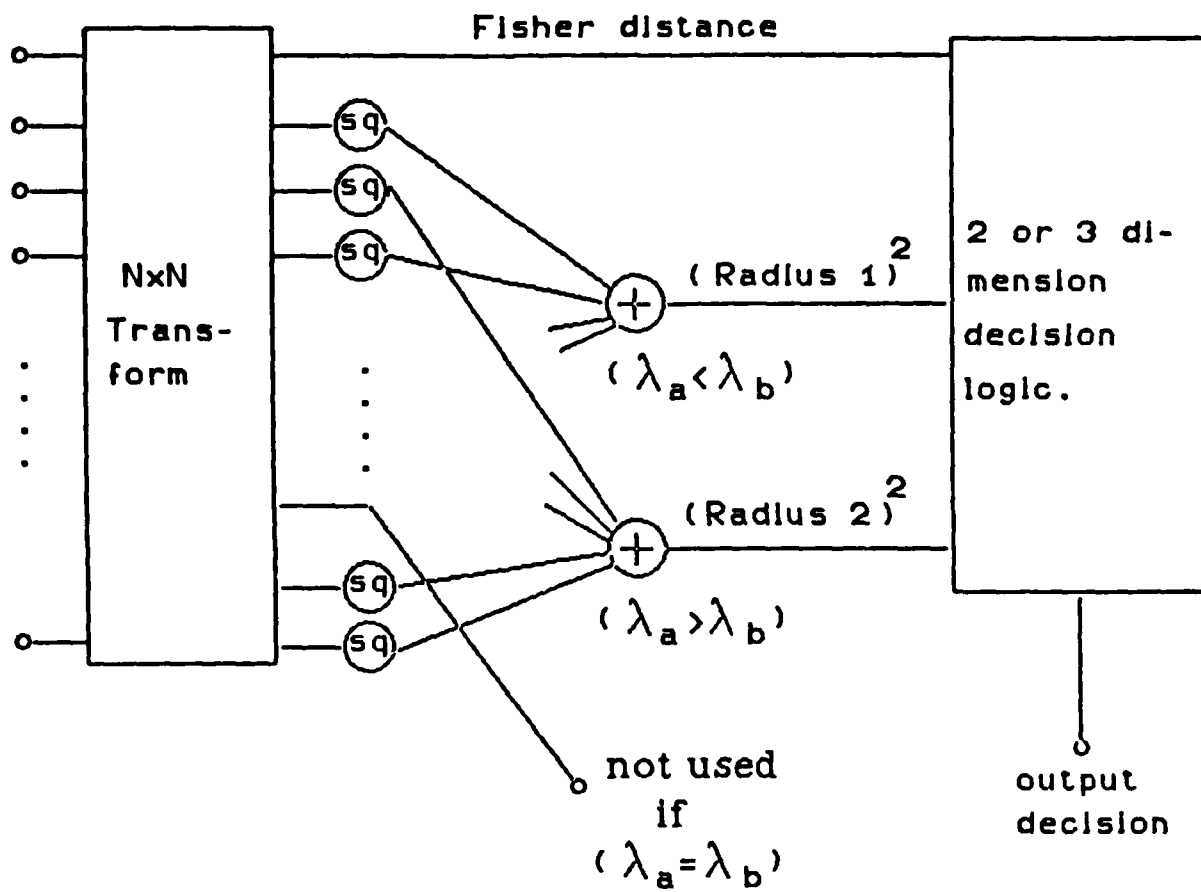$(\lambda_a = \lambda_b)$

2 or 3 di-
mension
decision
logic.

output
decision

Fig 10   Schematic of Fisher—Radius Classifier

DOCUMENT CONTROL SHEET

Overall security classification of sheet ..... UNCLASSIFIED ....

(As far as possible this sheet should contain only unclassified information. If it is necessary to enter classified information, the box concerned must be marked to indicate the classification eg (R) (C) or (S))

1. DRIC Reference (if known)	2. Originator's Reference Memorandum 3897	3. Agency Reference	4. Report Security Unclassified Classification
5. Originator's Code (if known)	6. Originator (Corporate Author) Name and Location Royal Signals and Radar Establishment		
5a. Sponsoring Agency's Code (if known)	6a. Sponsoring Agency (Contract Authority) Name and Location		

7. Title

ON EXTENSIONS TO FISHER'S LINEAR DISCRIMINANT FUNCTION

7a. Title in Foreign Language (in the case of translations)

7b. Presented at (for conference papers)   Title, place and date of conference

| 8. Author 1 Surname, initials
Longstaff   I D | 9(a) Author 2 | 9(b) Authors 3,4... | 10. Date | pp. ref. |
| 11. Contract Number | 12. Period | 13. Project | 14. Other Reference | |

15. Distribution statement

Unlimited

Descriptors (or keywords)

Dimensionality reduction, discriminant vectors, feature selection,

Fisher criterion, linear transformations, separability.

continue on separate piece of paper

Abstract
This Memorandum describes how Fisher's Linear Discriminant can be combined with the Fukunaga-Koontz transform to give a useful technique for reduction of feature space from many to two or three dimensions. Performance is seen to be superior in general to the Foley-Sammon extension to Fisher's method. The technique is the extended to show how a new radius vector (or pair of radius vectors) can be combined with Fisher's vector to produce a classifier with even more power of discrimination. Illustrations of the technique show that good discrimination can be obtained even if there is considerable overlap of classes in any single projection.

S80/48

END

FILMED

4-86

DTIC